DATA PAPER

Overview of SMP-CAIL2020-Argmine: The Interactive Argument-Pair Extraction in Judgement Document Challenge

Jian Yuan[1], Zhongyu Wei[1†], Yixu Gao[1], Wei Chen[1], Yun Song[2], Donghua Zhao[3], Jinglei Ma[4], Zhen Hu[4], Shaokun Zou[5], Donghai Li[6] & Xuanjing Huang[7]

[1]School of Data Science, Fudan University, Shanghai 200433, China

[2]Heilongjiang University, Heilongjiang 150080, China

[3]School of Mathematical Sciences, Fudan University, Shanghai 200433, China

[4]China Judicial Big Data Institute, Beijing 100043, China

[5]THUNISOFT Co., Ltd., Beijing 100084, China

[6]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[7]School of Computer Science, Fudan University, Shanghai 200433, China

Keywords: Argumentation mining; Judgement documents; Natural language understanding; Pretrained language model

Citation: Yuan, J., et al.: Overview of SMP-CAIL2020-Argmine: The interactive argument-pair extraction in judgement document challenge. Data Intelligence 3(2), 287-307 (2021). doi: 10.1162/dint_a_00094
Received: January 10, 2021; Revised: March 23, 2021; Accepted: April 25, 2021

ABSTRACT

In this paper we present the results of the Interactive Argument-Pair Extraction in Judgement Document Challenge held by both the Chinese AI and Law Challenge (CAIL) and the Chinese National Social Media Processing Conference (SMP), and introduce the related data set – SMP-CAIL2020-Argmine. The task challenged participants to choose the correct argument among five candidates proposed by the defense to refute or acknowledge the given argument made by the plaintiff, providing the full context recorded in the judgement documents of both parties. We received entries from 63 competing teams, 38 of which scored higher than the provided baseline model (BERT) in the first phase and entered the second phase. The best performing system in the two phases achieved accuracy of 0.856 and 0.905, respectively. In this paper, we will present the results of the competition and a summary of the systems, highlighting commonalities and innovations among participating systems. The SMP-CAIL2020-Argmine data set and baseline models[①] have been already released.

† Corresponding author: Zhongyu Wei (Email: zywei@fudan.edu.cn; ORCID: 0000-0003-3789-8507).

① https://www.disc.fudan.edu.cn/data/SMP-CAIL2020-Argmine

1. INTRODUCTION

In a trial process, the opinions, testimonies and results of both sides of the case are all recorded in detail in the judgement document [1], an example of which is shown in Figure 1. Traditionally, the summarisation of such text information remains to be organized and analyzed by the judge manually, which is highly time consuming and of low efficiency. In recent years, with the increasing interest in automatic analysis in the judicial field [2, 3, 4], more and more attention has been paid to an automatic system for judicial process, from Ulmer's proposal of quantitative methods and probability theory [5], Nagel's [6] optimization and statistical methods, to Liu & Chen's [7], Sulea et al.'s [8] and Katz et al.'s [9] natural language processing (NLP) models leveraging more lexical features in judicial documents, which indicates that such a task is greatly in need and of practical value.

Another research area of interest is argumentation mining, since argument is playing an increasingly important role in decision making on social issues. As an automatic technique to process and analyze arguments, computational argumentation, aimed at mining the semantic and logical structure of the given text, has become a rapidly growing field in natural language processing. Existing research on argumentation mining covers argument structure prediction [10, 11, 12], claims generation [13–17], and interactive argument pairs identification [18–24]. Recently, Cheng et al. [25] extracted argument pairs from peer review and rebuttal data in order to study the content, structure and the connections between them.

辩称：被告人包某甲对公诉机关指控的事实及罪名均无异议，不作辩解。
诉称：附带民事诉讼原告人吴某某及其委托代理人白1某诉称，要求依法追究被告人包某甲的刑事责任.....
裁判日期：二〇一五年十二月一日
本院认为：本院认为，被告人鳞甲因琐事故意伤害他人身体，致一人轻伤，其行为已构成故意伤害罪.....
审理经过：科右中旗人民检察院以右中检刑诉（201S）315号起诉书告人包某甲犯故意伤害罪，于201S年11月13日向.....
审判人员：审判长社某、审判员宋某、审判员白2某
公诉机关称：公201S年S月23日IS时许受害人吴某某、包某乙、除某某、白某某在新佳木苏木哈日巴达理查南山坡上放羊时.....
当事人：公诉机关科右中旗人民检察院。附带民事诉讼原告人吴某某，男，1989年10月3日出生，蒙古族，科右中旗人.....
本院查明：经审理查明，2015年5月23日1S时许，受害人吴某某、包某乙、陈某某、白某某在.....
裁判结果：一、被告人M甲犯故意伤害（轻伤）罪，判处拘役六个月；撤销201S年7月3日本院作出的.....
书记员：书记员王某

Figure 1. An instance of judgement document, which contains the statement of the defense and the plaintiff, the judgement date, the result of the trial, the judges' names, and the recorder's name.

In the works mentioned above, an interactive argument pair refers to the one that contains two arguments that have logical or semantic interactions with each other, e.g., "The global warming does not affect our daily life as the scientists say.", and "I cannot imagine what my life would be if my homeland is beneath the sea level.", which consists of two arguments, mainly talking about the same topic, the global warming in our examples, and the second one is responding to the first argument by hypothesizing the scene of global warming.

Since during the trial process, the two parties both have to make their own points clear and make response to the opposite party, which resembles the process of a debate to a large extent, and it is intuitive yet promising to apply computational argumentation methods to such a field. A typical task of this kind is to automatically extract the focus of dispute of the two parties in a trial process. Specifically, in a trail process, the focus of dispute between the plaintiff and the defense can refer to the arguments that two sides propose on fact statement or claim settlement, either consistent with each other or attacking each other, an example of which is shown in Figure 2, which is mainly the same with the setting of interactive argument pairs extraction. Therefore, such a task is of high practical value since the judge can be free from reading, comprehending, and analyzing the lengthy judgement documents manually with an automatic system to extract these focuses of dispute, and moreover, improve the efficiency and objectivity of the whole trial process.

No.	Party	Argument	Relationship
1.	诉称 (plantiff): 辩称 (defence):	2016 年 1 月 9 日，被告人谭某对谭某甲进行辱骂。 被告人谭某辩解称，其没有跟谭某甲发生过口角。	否认 (Denying)
2.	诉称 (plantiff): 辩称 (defence):	被告人谭某用原准备的木棒对其猛击，致其轻伤二级。 用于证实田某前后陈述有矛盾的地方。	否认 (Denying)
3.	诉称 (plantiff): 辩称 (defence):	现要求被告人谭某赔偿医疗费、法医鉴定费等费用，共计 18833.63 元。 对于民事赔偿部分，应当按过错比例进行分担。	部分自认 (Partially Acknowledging)

Figure 2. An example of three pairs of focus of dispute in one judgement document. Note: Each pair contains a sentence (i.e., argument) from the plaintiff and the defense, respectively. Among the three pairs, two of them are of Denying relationship and the other is of Partially Acknowledging relationship.

In order to address the aforementioned task, we hosted the Interactive Argument-Pair Extraction in Judgement Document (SMP-CAIL2020-Argmine) Challenge. We constructed a purpose-built data set that contains 4,080 entries of argument pairs from 976 judgement documents collected from http://wenshu.court.gov.cn/ published by the Supreme People's Court of China.

All the argument pairs are manually annotated by undergraduates and graduates majoring in law. Each of the argument pair consists of one argument from the plaintiff and the other from the defense that interacts with each other logically or semantically. During the process of annotation, annotators were given the full context of both sides and then required to extract all the interactive arguments between the plaintiff and the defense. Note that there can be multiple arguments from the defense that interact with the same argument from the plaintiff, and *vice versa*.

The task setting referred to the one designed in the Ji et al.'s work [23]. The systems participating in the SMP-CAIL2020-Argmine Challenge were required to identify the correct argument from the defense interacting with the given argument from the plaintiff among the five candidate arguments. That is to say, every entry of the collected argument pairs is converted into a multiple argument choice problem with four false options. Therefore, performing well in the task requires the system to deeply understand the semantic relationship of the given argument from the plaintiff and the candidate arguments. We conduct the competition in a two-phase fashion by setting threshold accuracy in the first phase, and only those whose system over-performs the baseline models we provide can enter the second phase. The number of argument

pairs reaches 4,080, including both the training data sets and the test data sets in two phases. In total, 315 teams from over 100 colleges and enterprises entered for the competition, 63 of which successfully submitted their models. We hope that research and practice in these fields will be stimulated by the challenges presented in this competition.

In this paper, we present a detailed description of the task and the data set, along with a summary of the submissions, and discuss the possible future research directions of the task.

2. RELATED WORK

2.1 Automatic Analysis of Judicial Documents

Automatic analysis of judicial documents has been studied for decades. At the very first stage, research tended to focus on mathematical and statistical analyses on existing court cases, instead of conclusions or methodologies on the prediction or summarisation of judicial documents. Ulmer proposed to suggest some uses of quantitative methods and probability theory in analyzing judicial materials [5]. Similar work including Nagel's [6] and Kort's [26] typically used optimization and statistics to conduct automatic judgement prediction. More recently, Lauderdale applied a kernel-weighted optimal classification estimator to recover estimates of judicial preferences [27].

These years have witnessed the booming in natural language processing (NLP), both theoretically and practically. As a natural application scenario of NLP, automation in judicial fields is also getting increasingly popular among NLP researchers. As a result, such automatic process of analyzing judicial documents has entered a brand new era. Liu and Chen [7] and Sulea et al. [8] extracted word features such as N-grams to train classifiers to predict the result of judgement, while Katz et al. [9] utilized case profile information (e.g., dates, terms, locations and case types). More advanced, Luo et al. introduced an attention-based neural model to predict charges of criminal cases, and verified the effectiveness of taking law articles into consideration [28].

Besides the automatic systems, a great number of interesting and meaningful tasks have also been proposed. For example, Xiao et al. [29] proposed a large-scale legal data set for judgement prediction, collected from China Judgments Online[②], and then organized a competition for this task [30]. After that, more judicial tasks and challenges were brought out such as Xiao et al. [31] and Liu et al. [32].

However, existing research mostly focuses on the case-level information understanding, such as applicable law articles, charges, and prison terms [29, 30], and insufficient research has noticed the importance of automatically extracting the focus of dispute, i.e., the interactive arguments from both sides of the case.

② http://wenshu.court.gov.cn/

2.2 Argumentation Mining

Argumentation mining is also a theoretical research area which has obtained much more attention, especially in the nearest years. As a research field in mining the logical and semantic structure in texts, various meaningful works have been proposed in recent years. For instance, Baff et al. [33] compared content- and style-oriented classifiers on editorials from the Liberal *New York Times* with ideology-specific effect annotations to explore the effect of writing style of editorials to audience of different parties; Ji et al. [23] proposed the task of identifying interactive argument pairs in online debate forum such as ChangeMyView (CMV), along with a novel representation learning method called Discrete Variational Encoder (DVAE) to encode different dimensions of information brought by the arguments in the corpus; Cheng et al. [25] collected the text data from peer review and rebuttal process to mine the argumentative relationship entailed in such discussion, and proposed a challenging data set of argument pair extraction with a multi-task learning framework to address such a task.

Also, the proposition of pretrained language models such as BERT [34] opens a brand new era of NLP, with impressively improved performance in nearly all tasks.

Obviously, the trial process greatly resembles the debate in many ways, since there are both two parties expressing their own opinions on the same topic and attacking each other's arguments. Therefore, it is practical to leverage models and methods in argumentation mining in the aforementioned judicial tasks.

3. DATA SET CONSTRUCTION

As discussed before, our goal is to construct an automatic system such that it can identify all the interactive argument pairs contained in the given judgement document which records the statement of both the plaintiff and the defense. Therefore, we collect the related data set from the judgement document corpus.

3.1 Data Source and Preprocessing

The raw data of judgement are provided by China Justice Big Data Institute, including over 10,000 entries in JSON format.

We first conducted random sampling on the raw data set, finding that there existed some documents of low quality. More specifically, the statement from the defense in some documents was so trivial, only containing the acknowledgement of all the statement made by the plaintiff; interaction of two sides in some documents only focused on the amount of charge, without any semantic or logical interactive arguments; and some documents contained too few or too many sentences to be analyzed.

In order to solve these problems, we refrained the data set with the following rules:

- Delete all the entries that contain "供认不讳" (forfeiting) or "无异议" (having no opposite opinions) in the first sentence of the defense's statement, since very few of these entries refute the statement of the plaintiff.
- Delete all the entries that contain less than two non-charging sentences in either statement of the plaintiff or the one of the defense (the "non-charging sentence" means the sentence that does not contain figures), as we do not hope the focus of dispute only aims at the amount of charge.
- Delete all the entries that contain less than four sentences in either statement of the plaintiff or the one of the defense, and all the entries that contain more than 1,500 words in the statement of both sides, so as to control the length of the data set, thus improving its quality.

After such filtering, we finally obtained 2,238 instances of judgement documents that are of high quality. Then we randomly sampled 40 of the obtained judgement documents and asked four graduate students to conduct human annotation of interactive argument pairs extraction. As a result, 120.25 argument pairs were extracted per person, and the average agreement was 0.628, which indicates that the task is both plausible and challenging.

3.2 Annotation

After preprocessing the raw data, we started the annotation of the data set. The platform used for annotation is shown in Figure 3, which acts as displaying the sentences in the judgement documents and saving the annotation results to database on the server.

We then employed six annotators who were undergraduates or graduates majoring in law, for more professional annotation. Each judgement document was annotated by two different annotators, in order to reduce the accidental error.

As shown in Figure 3, during annotation, the annotators were given the whole statement of both the plaintiff and the defense, with each sentence ordered and marked a number. Their task is then two-fold:

- Annotating features of the case. For the given case, annotators were required to specify some basic features of the whole case, including the case type, the type of the crime involved, as well as the entities of the plaintiff and the defense.
- Identifying all the interactive argument pairs in both sides' statement. The annotators then were required to identify all the interactive argument pairs entailed in the given case. Note that the amount of such pairs was not constant, so the annotators had to record all the interactive argument pairs by adding them one by one. Furthermore, we classified the argument pairs into four emotional categories: acknowledging, partially acknowledging, simple denying and active denying.

Figure 3. The online platform we used in the annotation, displaying the sentences in the judgement documents and saving the annotation results to the local server.

Note that in the second task, besides the identification of interactive argument pairs, the annotators were also required to classify each argument pairs collected. The four categories mentioned above represent different emotional polarities of the defense. Specifically, the argument pairs of acknowledging generally refer to the ones whose defense simply incorporates arguments like “I confess.”, partially acknowledging means the defense’s argument acknowledges some parts of the plaintiff’s but denying the others, simple denying contains the simple and direct denial such as “I did not hit the plaintiff.”, while the active denial

is more complicated, and sometimes it includes completely opposite statement on the same topic, e.g., "I did not hit the plaintiff, and instead, the plaintiff hit me with umbrella.". We conducted the classification for the purpose of making it more convenient for the judge to know which argument pairs needed further judgement and evidence. With these annotation standards, an instance of annotation is shown in Figure 4.

标注： 案件种类：民事案件 罪名/案由：虐待罪、遗弃罪 自诉人：张某 辩诉人：刘某
论点对及辩诉人态度：
被告人刘某在孩子张某出生后知道孩子有病不但不给治疗，还坚决把孩子扔到马路边，月子期间，被告人刘某不给孩子吃母乳，不让孩子喝水，不给孩子换尿布，孩子哭了也不哄。 > 被告人刘某辩称，1、自诉人张某既不是本案的受害人，也不是受害人的法定代理人，张某不能作为本案的自诉人起诉刘某；2、刘某从未实施过虐待、遗弃张某的行为。(单纯否认)
被告人刘某在孩子张某出生后知道孩子有病不但不给治疗，还坚决把孩子扔到马路边，月子期间，被告人刘某不给孩子吃母乳，不让孩子喝水，不给孩子换尿布，孩子哭了也不哄。 > 刘某在10个月的怀孕过程中，一直细心地照顾腹中的孩子，孩子出生后，张某及其儿子张某不管刘某，刘某及其父母整月都在细心地照料孩子张某的生活，从未虐待过孩子。(积极否认)
孩子满月那天，被告人刘某抛弃了孩子，偷走了为孩子治病的20000元手术费走了。 > 孩子满月时，张某及其家人强行将尚在怀抱中的张某抢走，并将刘某赶出家门。(积极否认)
六年多来，被告人刘某从没有看过孩子一眼，也没有为孩子治病拿过钱。 > 刘某曾多次回家看望张某，均遭到张某的阻拦。(积极否认)
六年多来，被告人刘某从没有看过孩子一眼，也没有为孩子治病拿过钱。 > 在张某出生后，刘某多次出钱为张某看病。(单纯否认)

Figure 4. The annotation result of No.12 judgement document, containing four interactive argument pairs extracted by the annotators.

3.3 Statistics on the Data Set

After six months of annotation, some basic statistics on the data set is shown in Table 1 below. From the table we can find that law major students indeed achieved higher agreement, indicating that professional knowledge helps improve the performance in this task. Another notable point lies in that interactive argument pairs, compared with all the sentence pairs in the corpus, are of very low density and bring challenges for automation.

Table 1. Basic statistics on the annotated data set.

Data set	Number
Annotated judgement documents	1,069
Annotated interactive argument pairs	4,476
Agreeable argument pairs	1,027
Disagreeable argument pairs	3,158
Sentence pairs in the annotated judgement documents	78,943
Average interactive argument pair density	0.058
Average agreement among annotators	0.960

4. TASK DESCRIPTION

4.1 Task Formulation

As mentioned above, the density of interactive argument pairs is very low (compared with all the sentence pairs between two sides), and thus we have to convert the identification task into an easier one. Our approach is to construct a multiple-choice problem for every argument from the plaintiff that occurs in at least one interactive pair, by adding four arguments from the defense that does not match the plaintiff's argument. That is to say, given an argument *sc* from the plaintiff, a candidate set of the defense's arguments consists of one positive reply bc^+, four negative arguments $bc_1^- \sim bc_4^-$, along with their corresponding contexts, and our goal is to automatically identify which argument from the defense has interactive relationship with the one from the plaintiff.

We formulated such a task as a 5-way multiple-choice problem. In practice, the participants' models calculated the matching score $S(sc, bc)$ for each argument in the candidate set with the plaintiff's argument *sc* and treated the one with the highest matching score as the winner. Note that here we did not use the emotional tags we collected before, since we would like to focus mainly on the identification of the correct argument pair in this competition.

Note that naturally, this setting needs the number of sentences in the statement of the defense to be no less than 5 (or more if there are not only one argument from the defense interacting with the plaintiff's one), so some of the entries are discarded and finally our whole data set comprises of 4,080 interactive argument pairs (i.e., multiple-choice problems) from 976 judgement documents. An example is displayed in Table 2 below.

Table 2. An example of the multiple-choice task.

Statement	Sentence
Full context of the plantiff	自诉人苏某某诉称：被告人康某的父亲马某与自诉人的母亲苏某原是夫妻……
Full context of the defense	被告人康某辩称：我的行为不构成故意伤害罪，……
The plantiff's argument	11时许，被告人康某骑一辆自行车来到现场，……
Candidate argument 1	她的伤是她自己雨伞造成的，……
Candidate argument 2	民事赔偿上，由于我没有对自诉人进行打击，……
Candidate argument 3	司法鉴定意见书不能作为本案的定案证据，……
Candidate argument 4	自诉人从事的是财务会计工作，……
Candidate argument 5	且鉴定时间和事故发生时间之间间隔了3个半月，……
Answer	**1**

4.2 Scoring Metric and Data Set Division

For the released multiple-choice task, we take accuracy as the evaluation metric. Specifically, if the ground truth of the *i*th problem is y_i, and the system predicts the answer to be $\hat{y}_i$, then the average accuracy on the test data set of size n is calculated as below:

$$\text{accuracy} = \frac{\sum_{i=1}^{n} y_i = \hat{y}_i}{n} \tag{1}$$

For the purpose of testing the system's generalization more fairly, we organized two phases in the competition and thus dividing the data set into three parts, namely SMP-CAIL2020-Argmine_train, SMP-CAIL2020-Argmine_test1, and SMP-CAIL2020-Argmine_test2. The quantity of these data sets is roughly 3:1:1.

In the first phase of the competition, participants were provided with the SMP-CAIL2020-Argmine_train data set to train their systems, and were tested with the SMP-CAIL2020-Argmine_test1 data set. Those who exceeded the performance of the given BERT baseline models were admitted to the second phase. And in the second phase, participants were provided with the SMP-CAIL2020-Argmine_test1 data set and tested with the SMP-CAIL2020-Argmine_test2 data set. The participants' final score = 0.3 * $Score_1$ + 0.7 * $Score_2$, in which the $Score_1$ and $Score_2$ means their score in two phases, respectively.

4.3 Baseline Models

Before we released the competition, we ran the following baseline models on the data set to obtain the border line for the admission to the second phase. Notice that for every baseline model, we only took the SMP-CAIL2020-Argmine_train data set as the training set.

- **All 1**
  This model directly output answer "1", which was used to examine whether the distribution of the answers was shuffled randomly enough.
- **Common Words**
  This model returned the candidate argument that had most common words with the given argument from the plaintiff, which was a simple and straightforward model leveraging lexical features.
- **BiLSTM**
  This model first conducted word segmentation using Jieba [35], and then we concatenated the plaintiff's argument with candidate arguments separately. In this way, we converted the 5-way multiple-choice into 5 sentence-pair classification problems. Then we randomly abandoned three negative sentence pairs so as to make the two classes balanced. For each sentence pair, their embedding was sequentially fed into a BiLSTM [36, 37] and took its final hidden state into a linear classifier to output the final prediction. The Figure 5(a) shows the model's overall framework.

- **BERT**
  BERT [34] is a pretrained language model based on transformers, and has proved to be exceedingly superior to many research aspects in NLP. In our experiment, we also converted the problem into the sentence-pair classification since it could be much easier to apply the BERT model to such a problem. The Figure 5(b) shows the model's overall framework.

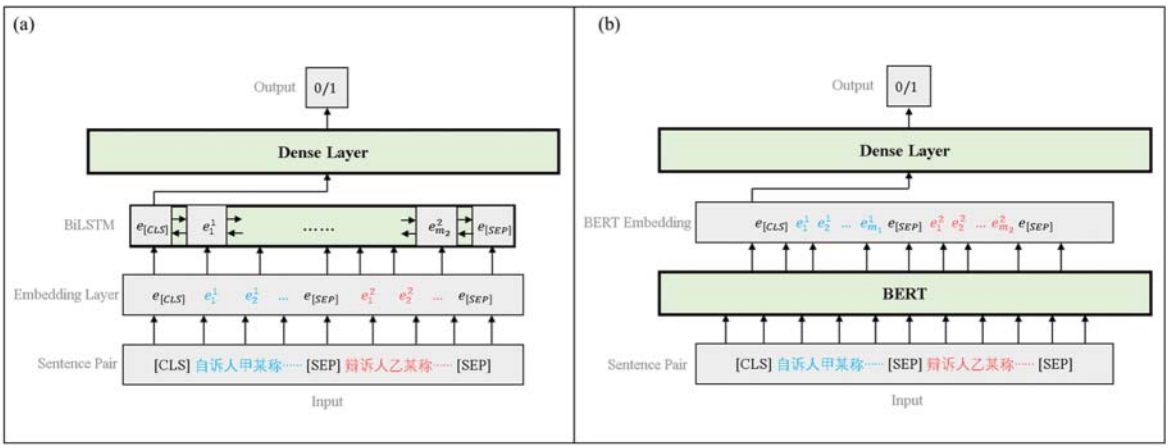


Figure 5. The overall framework of two neural network baseline models, in which (a) refers to the BiLSTM model and (b) refers to the BERT model.

All baseline models' performance is shown in Table 3 below. Since the best baseline model gives out an accuracy of 0.7476, we set the border line of the first phase at 0.75.

Table 3. Performance of all baseline models.

Model name	Train accuracy	Test1 accuracy	Test2 accuracy
All 1	0.2009	0.1890	0.1922
Common Words	0.4904	0.4908	0.5275
LSTM	0.8742	0.6270	0.6793
BERT	**0.8812**	**0.7476**	**0.7797**

4.4 Submissions

The SMP-CAIL2020-Argmine Challenge was hosted on CAIL[③], which allowed submissions to be scored against the blind test set without the need to publish the correct labels. The two phases of the scoring system were open from June 1 to July 9, and July 10 to August 3, 2020. Participants were limited to 3 submissions per week.

[③] http://cail.cipsc.org.cn

5. COMPETITION DETAILS

5.1 Participants and Results

There are over 300 teams from various universities as well as enterprises who have registered for SMP-CAIL2020-Argmine, 63 teams who have submitted their models in the first phase, and 21 teams who have submitted their final models. The final accuracy shows that neural models can achieve considerable results on the task, especially when given a larger training set. In Table 4, we list the scores of Top 7 participants of the task. We have collected the technical reports of these contestants. In the following parts, we summarize their methods and tricks according to these reports. The performance of all participants on SMP-CAIL2020-Argmine will be found in Appendix A.

Table 4. Performance of participants on SMP-CAIL2020-Argmine.

Team	$Score_1$	$Score_2$	Final score
zero_point	**0.852**	0.896	**0.8828**
a-U	0.816	0.901	0.8755
quanshuizhihuiguan	0.802	**0.905**	0.8741
i	0.811	0.886	0.8635
tiaodalanmao	0.800	0.857	0.8399
wf	0.788	0.853	0.8335
zhihuizhengfa	0.788	0.853	0.8335

5.2 The Submitted Models

5.2.1 General Architecture

Pretrained Language Model. Ever since BERT [34] was publicly proposed, the whole NLP area has been pushed into a new era, with almost all tasks improved in performance. Also, among the baseline models above, BERT gives out the best performance on the task, and therefore makes the pretrained language model such as Sentence-BERT [38], RoBerta [39], and ERNIE [40] popular in submissions.

Fine-tuning Mechanisms. After leveraging the pretrained models mentioned above to obtain embedding for tokens and sentences, fine-tuning is needed to further improve the model's performance, including:

- **Attention.** A natural idea to further fine-tune the representation of the arguments is to leverage the attention mechanism between the plaintiff's argument and five candidate arguments separately.
- **RNN Layers.** Note that after using the pretrained models, we have token-level, sentence-level as well as sentence-pair-level representation (the representation of [CLS]). Therefore, we can retain the sentence-pair-level representation, and feed the tokens' embedding into another BiLSTM layer and concatenate them before the linear classifier.
- **Memory Networks.** All the methods mentioned above only use the information of the arguments. However, we have provided the whole context of both sides in the judgement documents. Hence, it is plausible to use memory networks [41] to retrieve the context information.

5.2.2 Promising Tricks

Other than the standard "pretrained model + fine-tuning" mode, there are some useful tricks which can address the issues met in the task and improve the sentence pair classification models significantly. We summarize them as follows:

Fine-tuning with external corpus. Teams such as "zero_point", "quanshuizhihuiguan" as well as "tiaodalanmao" all tried to fine-tune their pretrained model by adding external judicial corpus. Such a method helps improve the model since external judicial corpus enables the pretrained language models to learn more topic-specific language knowledge and therefore performs better in judicial settings. As is reported by them, this method enables the model to have an increase in accuracy by about 1%.

Data Augmentation and Data Balancing. The "a-U" team followed our way of constructing the multiple choices and generated more multiple-choice questions for training by retrieving more negative samples from the provided contexts of the defense, which helps the model to further leverage the context information and incorporate more textual knowledge. Moreover, to address the problem of data imbalance (too many negative samples), they used over-sampling on positive instances to avoid the model's getting lost in the overwhelming size of negative samples.

Loss Function. Most models use cross entropy as their loss functions. However, some models adopt more promising loss functions, such as focal loss [42] to enhance the performance on low frequency categories, and triplet loss to improve the model's ability of generalization. Besides, the loss weights of various categories and the activation functions of the output layer also have great influence on the final performance. As is reported by the competitors, such a method transforms the task into an argument pair ranking problem, instead of the classification problem, which helps the model to gain an improvement of over 4%.

Model Ensembling. Some participants trained several different classification models over different samples from the whole data set, and finally combined them with majority voting or weighted average strategies to combine their predicting results. Among all the participants using such a method, the "a-U" team trained five sequence classification models based on BERT and adopted the majority voting method to reduce the variance of a single model, therefore improving the robustness of the model, which finally helps their model to achieve the second prize of the competition.

5.2.3 Error Analysis

Here, we inspect the erroneous outputs of our model to identify major causes of mismatches. There are mainly two issues.

Sentence Length Limitation in Pretrained Models. Since pretrained models like BERT have maximal length limitation, i.e., they will truncate sentence pairs that contain huge size, thus making the model unable to process all the information entailed in the sentence pair.

Entity Mismatch. Among many false cases, the error caused by entity mismatch is quite common. In the cases where there are multiple defences, the plaintiff may propose different prosecutions to different defences. However, some of them may share the same action mentioned by plaintiff, thus making the model confused when the negative candidate argument contains the detailed action while the positive one only includes simple denial.

6. CONCLUSION AND FUTURE WORK

In SMP-CAIL2020-Argmine, we employ the interactive argument-pair extraction in judgement document as the competition topic. In this competition, we construct and release a brand new data set for extracting the focus of dispute in the judgement documents. The performance on the task was significantly raised with the efforts of over 300 participants. In this paper, we summarize the general architecture and promising tricks they employed, which are expected to benefit further research on legal intelligence. However, there is still a long way to go to fully achieve the goal of automatically extracting the focus of dispute since the task is already a simplified one. Also, leveraging some more case-based features such as the type of case and type of crime and the semantic label of the interactive argument pairs may possibly further improve the model's performance.

ACKNOWLEDGEMENTS

This work is partially supported by National Key Research and Development Plan (No. 2018YFC0830600), and is cooperated with China Justice Big Data Institute, which provided judgement documents and the employment of professional annotators. The competition is also sponsored by Beijing Thunisoft Information Technology Co., Ltd., and supported by both CAIL and SMP organizers.

AUTHOR CONTRIBUTIONS

All of the authors have made meaningful and valuable contributions to the resulting manuscript. J. Yuan (19210980107@fudan.edu.cn) undertook the code running test of the task, summarized the evaluation task and drafted the paper. Y. Gao (yxgao19@fudan.edu.cn) and W. Chen (chenwei18@fudan.edu.cn) participated in providing baseline models to the contestants. Z. Wei (zywei@fudan.edu.cn), S. Zou, D. Li (lidh18@mails.tsinghua.edu.cn), D. Zhao (dhzhao@fudan.edu.cn) and X. Huang (xjhuang@fudan.edu.cn) designed, released and promoted the shared task. Y. Song (1171991@s.hlju.edu.cn), J. Ma (mqstssf2009@126.com) and Z. Hu (huz06@126.com) helped formulate the shared task from a professional law perspective.

DATA AVAILABILITY STATEMENT

The data sets generated and analyzed in the study are not currently available to the public due to the fact that the data sets are produced by judicial expert consultants of China Judicial Big Data Institute based on their professional knowledge and experience. The publicly released version of the data sets needs the consent of all expert consultants, and hence currently it can only be accessed from the corresponding author on reasonable request.

REFERENCES

[1] Vermeule, A.: Judicial history. Yale Law Journal 108, 1311 (1998)
[2] Long, S., et al.: Automatic judgment prediction via legal reading comprehension. In: China National Conference on Chinese Computational Linguistics, pp. 558–572 (2019)
[3] Segal, J.A.: Predicting Supreme Court cases probabilistically: The search and seizure cases, 1962–1981. American Political Science Review 78(4), 891–900 (1984)
[4] Keown, R.: Mathematical models for legal prediction. Computer/I J 2, 829 (1980)
[5] Ulmer, S.S.: Quantitative analysis of judicial processes: Some practical and theoretical applications. Law and Contemporary Problems 28(1), 164–184 (1963)
[6] Nagel, S.S.: Applying correlation analysis to case prediction. Texas Law Review 42 (1963)
[7] Liu, Y.-H., Chen, Y.-L.: A two-phase sentiment analysis approach for judgement prediction. Journal of Information Science 44(5), 594–607 (2018)
[8] Sulea, O.-M., et al.: Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306 (2017)
[9] Katz, D.M., Bommarito, M.J., Blackman, J.: A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE 12(4), e0174698 (2017)
[10] Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 46–56 (2014)
[11] Liu, J., Cohen, S.B., Lapata, M.: Discourse representation parsing for sentences and documents. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6248–6262 (2019)
[12] Wang, L., et al.: Predicting thread discourse structure over technical web forums. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 13–25 (2011)
[13] Bilu, Y., Slonim, N.: Claim synthesis via predicate recycling. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 525–530 (2016)
[14] Zukerman, I., McConachy, R., George, S.: Using argumentation strategies in automated argument generation. In: INLG'2000 Proceedings of the First International Conference on Natural Language Generation, pp. 55–62 (2000)
[15] Sato, M., et al.: End-to-end argument generation system in debating. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, pp. 109–114 (2015)
[16] Hua, X., Wang, L.: Neural argument generation augmented with externally retrieved evidence. arXiv preprint arXiv:1805.10254 (2018)
[17] Zhao, T., Lee, K., Eskenazi, M.: Unsupervised discrete sentence representation learning for interpretable neural dialog generation. arXiv preprint arXiv:1804.08069 (2018)
[18] Taghipour, K., Hwee, T.N.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891 (2016)
[19] Wei, Z., Liu, Y., Li, Y.: Is this post persuasive? Ranking argumentative comments in online forum. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 195–200 (2016)
[20] Tan, C., et al.: Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In: Proceedings of the 25th International Conference on World Wide Web, pp. 613–624 (2016)
[21] Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162 (2017)
[22] Habernal, I., Gurevych, I.: Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1589–1599 (2016)

[23] Ji, L., et al.: Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3703–3714 (2018)
[24] Ji, L., et al.: Discrete argument representation learning for interactive argument pair identification. arXiv preprint arXiv:1911.01621 (2019)
[25] Cheng, L., et al.: Argument pair extraction from peer review and rebuttal via multi-task learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7000–7011 (2020)
[26] Kort, F.: Predicting Supreme Court decisions mathematically: A quantitative analysis of the "right to counsel" cases. The American Political Science Review 51(1), 1–12 (1957)
[27] Lauderdale, B.E., Clark, T.S.: The Supreme Court's many median justices. American Political Science Review 106(4), 847–866 (2012)
[28] Luo, B., et al.: Learning to predict charges for criminal cases with legal basis. arXiv preprint arXiv:1707.09168 (2017)
[29] Xiao, C., et al.: CAIL2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
[30] Zhong, H., et al.: Overview of CAIL2018: Legal judgment prediction competition. arXiv preprint arXiv:1810.05851 (2018)
[31] Xiao, C., et al.: CAIL2019-SCM: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962 (2019)
[32] Liu, C.-L., Hsieh, C.-D.: Exploring phrase-based classification of judicial documents for criminal charges in Chinese. In: International Symposium on Methodologies for Intelligent Systems, pp. 681–690 (2006)
[33] Baff, R.EI., et al.: Analyzing the persuasive effect of style in news editorial argumentation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3154–3160 (2020)
[34] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
[35] Sun, J.: Jieba Chinese word segmentation tool. Available at: https://github. com/fxsjy/jieba. Accessed 25 June 2018
[36] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
[37] Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. In: The Ninth International Conference on Artificial Neural Networks ICANN 99, pp. 850–855 (1999)
[38] Reimers, N., Iryna, G.: Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084 (2019)
[39] Liu, Y., et al.: RoBerta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
[40] Zhang, Z., et al.: ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)
[41] Sukhbaatar, S., Weston, J., Fergus, R.: End-to-end memory networks. In: Advances in Neural Information Processing Systems, pp. 2440–2448 (2015)
[42] Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

APPENDIX A: FULL RANK OF ALL PARTICIPANTS

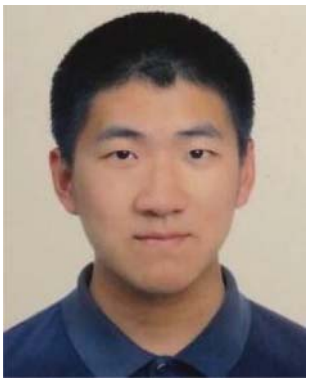
Full rank of all participants in CAIL-SMP2020-Argmine. $Score_1$ and $Score_2$ refer to the score achieved by the participants in phase I and II, respectively, while Final Score refers to the weighted sum of $Score_1$ and $Score_2$.

Table A1. Full rank of all participants.

Team	$Score_1$	$Score_2$	Final Score
zero_point	0.852	0.896	0.8828
a-U	0.816	0.901	0.8755
quanshuizhihuiguan	0.802	0.905	0.8741
i	0.811	0.886	0.8635
tiaodalanmao	0.800	0.857	0.8399
wf	0.788	0.853	0.8335
zhihuizhengfa	0.788	0.853	0.8335
quanzhizhixing	0.789	0.852	0.8331
bl_ssk	0.787	0.852	0.8325
xiaocuiwawa	0.785	0.852	0.8319
fabaozhineng	0.796	0.847	0.8317
fajixianzonghewozuodui	0.785	0.851	0.8312
zhuimengzhizixin	0.794	0.847	0.8311
CBD	0.779	0.853	0.8308
xiaofa	0.777	0.852	0.8295
xiaozhineng	0.775	0.840	0.8205
testing	0.756	0.845	0.8183
falvzhineng	0.763	0.841	0.8176
boys	0.760	0.826	0.8062
301deshuishou	0.768	0.810	0.7974
sos	0.755	0.797	0.7844
qilejingtu	0.780		
TEEMO	0.780		
qweasd	0.774		
maitianxback	0.772		
duimingmeixianghao	0.771		
wisdom	0.768		
anonymous	0.768		
seu	0.768		
DN	0.768		
hongseyoujiaosanbeisu	0.768		
ooo	0.768		
OO	0.768		
zhangyuanyu	0.768		
daminghu	0.757		
zunjisoufa	0.757		
yunshujingjixue	0.755		
DL	0.753		
tiantianxiangshang	0.751		

Team	$Score_1$	$Score_2$	Final Score
jizhikekeyupipi	0.751		
ddlqianzuihouchongci	0.750		
Tracee	0.748		
zhineng	0.744		
heitu	0.736		
chong!	0.728		
nlpxiaoxuesheng	0.725		
sr	0.719		
huangjinkuanggong	0.714		
hello	0.708		
aaaa	0.706		
houchangcunbaoan	0.704		
EC_lab	0.680		
imiss	0.672		
nnnn01	0.629		
zhegexiaohaiyoudiandou	0.598		
woshijiangdaqiao	0.520		
nlp11	0.517		
mushangdaren	0.491		
Eupho	0.491		
LawBoys	0.491		
xuexijishudui	0.491		
test11	0.472		
lw	0.344		
amazing	0.083		

AUTHOR BIOGRAPHY



Jian Yuan is currently a graduate student of the School of Data Science, Fudan University. His research interests include argumentation mining, legal artificial intelligence and knowledge representation.
ORCID: 0000-0002-3201-9844

Zhongyu Wei is an Associate Professor in School of Data Science at Fudan University and he serves as the secretory in Social Media Processing (SMP) committee of Chinese Information Processing Society of China (CIPS). He got his PhD in The Chinese University of Hong Kong in 2014. His research focuses on multi-modality information understanding and generation cross vision and language, argumentation mining and some cross-disciplinary topics.
ORCID: 0000-0003-3789-8507

Yixu Gao is currently a graduate student in the School of Data Science, Fudan University. Her research interests include argument mining, reinforcement learning and recommendations.
ORCID: 0000-0002-6605-6416

Wei Chen is currently a PhD student in the School of Data Science at Fudan University. His research interests include dialogue systems and natural language generation.
ORCID: 0000-0001-9431-9247

Yun Song is a doctoral student of the Law School, Heilongjiang University, Harbin, China. Her research interest includes the history of law, artificial intelligence and justice.
ORCID: 0000-0001-5032-0107

Donghua Zhao is an Associate Professor of the School of Mathematical Sciences, Fudan University, Shanghai, China. She received her PhD degree in Applied Mathematics from Fudan University in 2005. Her research interest includes differential equations, complex networks, natural language processing and time-series analysis.
ORCID: 0000-0002-4959-2647

Jinglei Ma is a PM (product manager) of the China Judicial Big Data Institute. He graduated from the Beihang University with a Master of Laws. His research interest is judicial big data.
ORCID: 0000-0001-5854-2425

Zhen Hu is an independent researcher who is interested in machine learning, natural language processing and control system. He received his PhD degree in Automation from Tsinghua University in 2015, and led some projects about smart city, legal Intelligence and some other related subjects afterward.
ORCID: 0000-0001-9587-3493

Shaokun Zou is the Chief Executive Officer (CEO) of Beijing Huayu Yuandian Information Service Co., Ltd. His research interest involves legal artificial intelligence and automatic legal systems.

Donghai Li is a professorate senior engineer, deputy general manager and chief technology officer of Beijing Huayu Yuandian Information Service Co., Ltd. He is a D.Eng. candidate in the Leading Talents for Innovation Program in Department of Computer Science and Technology in Tsinghua University. His major research interests are in legal search technologies. He is dedicated to research and application of legal artificial intelligence, one of the first few to apply knowledge graph technologies to the law tech field.
ORCID: 0000-0001-5177-3335

Xuanjing Huang is a Professor of the School of Computer Science, Fudan University, Shanghai, China. She received her PhD degree in Computer Science from Fudan University in 1998. Her research interest includes artificial intelligence, natural language processing, information retrieval and social media processing.
ORCID: 0000-0001-9197-9426